



EVALUATION OF WEST AFRICAN EXAMINATIONS COUNCIL'S PHYSICS ESSAY TEST USING PARTIAL CREDIT ITEM RESPONSE THEORY MODEL

¹Dr. Fidelis O. Nnadi & ²Dr. Rose C. Anamezie

^{1&2} Department of Science Education,
Enugu State University of Science and Technology (ESUT), Enugu-Nigeria

ABSTRACT

The poor state of secondary school students' achievement relative to policy expectation in Physics triggered the study. At the level of instrument quality, the scoring pattern in West African Senior School Certificate Examination's (WASSCE) Physics essay marking scheme, difficulty-parameter level of the items and differential item functioning were sought. Instrumentation research design was adopted. One hundred and sixty senior secondary III Physics students consisting of 63 female and 97 male students formed the sample. The 2018 WASSCE Physics essay with its marking scheme formed the instrument used to collect data. The internal consistency reliability of the instrument sought using Kuder-Richardson's formula 21 had value of 0.81. Three research questions and one null hypothesis guided the study. Research questions were answered using frequencies, mean infit/outfit statistic, mean item difficulty and mean standard error. The null hypothesis was tested at .05 level of significance using Mantel-Haenszel's Chi-square probability value. The result indicated that fractional-polarization scoring was adopted for two items. 25 in 28 items fit the partial credit model. The items were moderately difficult and did not exhibit gender-based differential item functioning. It is recommended that West African Examinations Council should remove fractional-dichotomization scoring and breakup marks allocated to an item into more than two groups for all items in future Physics essay marking scheme.

Keywords and Phrases: Item-response-theory, partial-credit-model, fractional-dichotomization, scoring and Physics essay.

INTRODUCTION

The development of any nation is anchored on the level of scientific and technological (S & T) literacy possessed by its citizens. Science Education field, no doubt aims at producing a nation's citizens that can adapt to a new world-order of science and technological innovations. S & T innovations are aimed at solving myriads of problems posed to humans. It is understandable that in the cause of solving human problems following the rudiments of science and technology, more problems can sometimes emerge. This is chiefly because scientific and technological processes are man-made and as such they are error-laden. The present day science and technology aim at reduction of errors. Despite the short-comings



of science and technology, the society has witnessed unprecedented increase in the quality of life occasioned by S & T discoveries. Due to the importance of S & T discoveries, any developed economy of the world uses them as indicators of world power (Tolu & Olusegun, 2014). In realization of the importance of science for a nation's global economic and political competitiveness, the federal government of Nigeria (FGN, 2013) through the National policy on Education included Physics as one of the core science subjects to be studied in upper basic schools. In the same statutory document, FGN also reiterated the relevance of Science to Nigerian economy by recommending 60:40 admission ratio into tertiary levels of education in favour of science and its allied fields. The essence of the admission ratio is to produce more personnel in the Sciences relative to Arts, who will be able to drive the nation's economy with the requisite skills needed in the new age of Science and Technology. It is quite disheartening to note that secondary school students' achievement in Physics, one of the science fields especially in external examinations has been reported poor (Telima, 2018). The poor state of achievement in Physics has constituted a source of worry to researchers. So many factors have been linked to the poor achievement in Physics, including the use of instrument of poor quality during the testing processes (Nworgu, 2016). Poor instrument quality can manifest in the form of poor scoring pattern adopted in the marking scheme. For example, it is a common practice among teachers and examination bodies to award dichotomous scores to an essay test scored with partial credit. Sometimes the dichotomous scores are fractionalized, for example "award zero-mark or 2.5marks". Partial credit scoring demands that any score awarded to an essay item should be split into many parts, preferably more than two parts. Items of poor validity and reliability cannot be used to ascertain the true ability of the test-takers. The indices of validity of essay questions determined through psychometric forensics also included difficulty, discriminatory power, carelessness, item characteristic curve and differential item functioning (DIF). A good essay item should neither be too difficult nor too easy. It should be able to discriminate between upper and lower ability-level students. A good essay test should be able to measure amount of carelessness exhibited by the students. The shape of the item characteristic curve for a good item should observe a logistic or probit regression curve: "s-shaped" A good essay test item should also not favour a particular sub-group (focal) and disfavor another sub-group (reference) of same ability.

The West African Examinations Council (WAEC) is an examination body that is saddled with the responsibility of conducting valid and reliable examinations across West African sub-region. Nigeria as a member of league of five West African states including Ghana, The Gambia, Sierra Leone and Liberia established WAEC in 1952 to enhance the quality of secondary education in the country, in terms of developing and validating objective questions, essay and practical skills in relevant subjects, in addition to certification. WAEC is the premier examination body in the West African sub-region. For easy administration, WAEC placed Nigeria and The Gambia into one group. Thus, Nigeria and The Gambia have uniform curriculum, questions and marking scheme in all school



subjects including Physics. WAEC has been reported to adopt the Item Response Theory (IRT) in evaluating its tests. The IRT is a cutting-edge psychometric model used for item-analyses of both dichotomously-scored (DS) and polytomously-scored (PS) responses. IRT is a family of probabilistic models which describes the linkage between examinees' ability and performance in a set of items (Sung & Kang, 2006). One of the IRT models for evaluation of an essay test is the partial credit model. IRT models are invoked as a result of the shortcomings of classical test theory (CTT), which hitherto has been in use for determining psychometric qualities of test items by WAEC. Nworgu (2010) indicated that IRT was a better alternative to CTT in test development process. CTT posits that an examinee's score is a function of the true score plus an error term. In addition, estimates of item and person parameters are dependent on the sample size used, not the latent construct (ability) of the respondents being measured (Matteucci & Stracqualursi, 2006). In addition, Ugodulunwa (2015), in support of sample-size dependent result of CTT also indicated that the meaning of any score an examinee had was dependent on the group the examinee belonged. Ugodulunwa further stressed that the difficulty of an item for instance depended on the group it was administered, not the ability being measured by the items. The challenge posed by sample-size-dependent result of CTT is in terms of making generalizations to the inaccessible target population. Another drawback of CTT is that both item and person-parameters are jointly analyzed. Item and person-oriented parameters are not separated in the CTT framework. CTT treats both item and person-parameters as discrete variables

Under the IRT framework, estimation of item and person-oriented parameters are independent of sample size used. Rather, estimation of parameters depends on the latent ability of the test-takers being measured. Item and person parameters are estimated in such a way that they are mutually exclusive. Estimation of one parameter does not influence the other. Most importantly, IRT conceptualizes item and person parameters as stochastic variables. One important feature of IRT is that when test data meet model assumptions, item parameter estimates are invariant over different samples drawn from the population (Qingping, 2010). Invariance of the values of parameter estimates under IRT implies that they are equivalent across different samples. Qingpin also noted that the use of IRT for analyzing test was predicated on some assumptions that should be met by the data. Items that violated the IRT assumptions were usually recast or sent to content experts for their opinions before they were processed for item banking, test equating, computerized adaptive testing or remote proctoring.

For test items that measure a single latent construct, unidimensionality assumption holds for such items. A test is unidimensional when a single latent trait accounts for all the common variance among its items (Thorpe & Favia, 2012). In other words, a unidimensional test has all the variances of its eigen-vectors (representing data) loaded on a single latent construct which can be termed content area after exploratory factor analysis. Violation of the assumption of unidimensionality occurs when the variances of a test's eigen-vectors are loaded on more than one latent construct. Violation of the usage



assumption of unidimensional test implies that the result of such test is sample dependent. However, violation of the unidimensionality assumption can only be allowed for tests that are designed to measure multi-dimensions or skills. Tests with multi-dimensional assumptions can be analyzed using multi-dimensional item response theory (MIRT).

Another usage assumption of IRT is monotonicity. Monotonicity of items upholds that item endorsement should increase as latent trait or ability level of the respondents increase (Thorpe & Favia, 2012). Monotonicity of items can be established when a high point-biserial correlation exists between an item and test score. Visual inspection of a test's monotonicity is also achieved using a test's information curve. The presence of a logistic or probit graph which is usually in the shape of letter "S" indicates monotonicity of test items. Qingpin (2010) advised that test takers ability should increase with probability of success in an item, provided that the items difficulty levels are constant. Constant item difficulty can be achieved when there is no meaningful variation between a focal (male) and reference (female) sub-groups' difficulty levels.

Local independence assumption of IRT upholds that there exists a weak positive correlation between the items of a test. The presence of positively high correlation between test items suggests similarity. Anamezie and Nnadi (2018) indicated that local independence implied that the probability of an examinee's correct scoring of an item did not influence correct or incorrect scoring of another item. Adonu (2014) also indicated that locally independent test items were mutually exclusive.

The partial credit model, PCM (Masters, 1982) is one of the polytomous models of IRT family of models. It evolved from the generalized partial credit model, GPCM (Muraki, 1992). In the PCM according to Primi and Primi (2014) the probability, P_{nix} of student's ability, θ_n is related to item difficulty parameter, b_{ij} and discrimination parameter, α_i as follows:

$$P_{nix}(\theta_n) = \frac{\exp \sum_{j=1}^x \alpha_i [\theta_j - (\theta_n - b_{ij})]}{1 + \sum_{k=1}^m \exp \sum_{j=1}^k (\theta_n - b_{ij})}$$
 indicates the probability of subject n obtaining a score of x on item i (item scores have a notation $x = 0, \dots, m$ where $m + 1$ equal to the number of response categories); parameter θ_n is the measure of person n in the latent dimension θ measured by the item; $\text{Exp}(x)$ is the symbol referring to natural exponential function $e \cong 2.72$ to the power x ; and parameter b_{ij} is the point of transition or threshold between the scores of item i . (p.6)

The PCM is an extension of dichotomous type of Rasch one-parameter logistic model. PCM models only the difficulty-parameter b_{ij} by constraining the discrimination-parameter α_i to a value of one. The PCM is efficient in analyzing data that are partially scored. Partial-scoring for an examinee j , under item i , category x refers to a system of scoring where marks are awarded unevenly across the parts of an item in a test. When marks are awarded in partial credit, there should be no polarization of the marks to be awarded for an item, otherwise the internal validity of such item will be unstable. Polarization of marks in partial



credit scoring entails for example awarding zero or one mark, zero or two marks and vice versa for an essay-type item. The difficult parameter is only estimated keeping discrimination parameter constrained to a value of one in PCM. PCM is chosen because the researchers' interests centered on ascertaining the difficulty level of West African Senior School Certificate Examination (WASSCE), occasioned by consistent poor achievement in Physics.

There is dearth of studies on essay test evaluation using IRT. Prama, Supahar, Warsono and Jumadi (2018) investigated a fifteen Physics essay question using partial credit model. Infit mean square and item difficulty values were used to answer the research questions posed in the study. Quest and parscale computer programs were used for the analysis. The result indicated that 13 in 15 items with infit mean square values ranging from .80 to 1.18 fit the model. The difficulty index of the items ranged between -1.15 to 1.58. Also, Adonu (2014) investigated the psychometric qualities of WAEC and NECO 2011 and 2012 practical Physics using winsteps computer programme. Part of the result indicated that the standard error of measurement of the test was low for all items: below .18. The fit statistic measured using infit and outfit mean square indicated that 22 in 24 items in 2011 and 2012 fit the model respectively. The items exhibited moderate item difficulty with values ranging from -1.53 to 1.94. However, there has been paucity of research evidences on the credibility of WAEC Physics essay test. The problem of this study therefore is to determine the validity of WAEC Physics essay questions using partial credit model.

PURPOSE OF THE STUDY

The study was designed to:

- (1) Determine the scoring pattern adopted by WAEC in 2018 WASSCE Physics essay test marking scheme for Nigeria and The Gambia.
- (2) Determine the proportion of item categories of 2018 WASSCE Physics essay test that fits partial credit model.
- (3) Determine the values of difficulty-parameter with standard error of measurement of the test item categories.
- (4) Determine the differential distractor functioning of the test items with respect to students' gender.

RESEARCH QUESTIONS

The following research questions guided the study:

- (1) What pattern of scoring was adopted by WAEC in 2018 WASSCE Physics essay test marking scheme for Nigeria and The Gambia?
- (2) What proportion of item categories of 2018 WASSCE Physics essay test fits partial credit model?



- (3) What are the values of difficulty-parameter with standard error of measurement of the test item categories?

HYPOTHESIS

The study was anchored on one null hypothesis:

H₀₁: The item categories of 2018 WASSCE-Physics-Essay test do not exhibit uniform differential item functioning across students' gender.

METHOD

The research design adopted was instrumentation. A sample of one hundred and sixty senior secondary three (SS3) Physics students consisting of ninety-seven male and sixty-three female students nested within four intact classes from Queen of the Rosary secondary school and St. Teresa's college in Nsukka education zone of Enugu state, Nigeria was used. Balloting with replacement was used to sample the schools and SS3 science class in a school with many streams of SS3 science students. Three instruments used to collect data included (i) prototype of year 2018 WASSCE Physics essay questions with spaces for answers (ii) year 2018 final marking scheme in Physics essay for Nigeria and The Gambia (iii) Differential item functioning, DIF checklist developed by the researchers. The 2018 Physics essay questions were developed and found valid by West Africans Examinations Council (WAEC) consisted of two parts: A and B. Part A provided demographic information of the test takers. Part B consisted of 28 essay items (1a-12d) covering ordinary level Physics curriculum. The marking scheme for the Physics essay consisted of marks distribution for a 28-item essay test. Table 1 shows the summary of marks allocated to each question. The same pattern of scoring was adopted in the study. The DIF-checklist was a 28-item by 160 respondents' binary matrix of zero "0" and one "1". Score '1' indicated students' item mastery level from 60% and above while '0' indicated students' item mastery level below 60%. A student who scored '1' was presumed to find the item easy, while a student who scored '0' was presumed to find the item difficult. The researchers determined the binary matrix from student's scores in each essay item. The Physics teachers in their respective schools distributed the questions under close watch by one of the researchers. The students were asked to respond to all items. Controlling for *speededness*, a situation an item difficulty level was affected by testing time, the research subjects were allowed to submit their answer sheets when they finished the test. Twenty-two copies of the Physics essay-test questions were administered on parallel sample in Enugu Education. Kuder-Richardson's formula-21 yielded an alpha value of .87 on the data generated. The result indicated a high internal consistency reliability for items scored with partial credit.

For preliminary analyses, the assumptions for use of item response theory models were tested. Firstly, monotonicity of items was established by conducting point-biserial correlation between each binary item responses and the continuous total test score using



latent trait model (ltm) package and specifically biserial.cor function in R version 3.4.3 computer programme. A coefficient whose value ranged between .04 to .46 with probability values for each correlation pair above .05 indicated high point-biserial correlation and presence of monotonicity and discrimination (from classical point of view) of the test items. Secondly, unidimensionality assumption was tested using exploratory factor analysis (EFA) to determine the number of latent constructs the test items were attached to. Specifically, principal component method of factoring was adopted because it had the propensity to yield more factors that loaded heavily on the construct. The extraction method was based on eigen- values greater than one. The rotation method applied was varimax, because it maximized the sum of the variances of squared loadings. Any coefficient below .35 was suppressed. The result of EFA indicated that the twenty-eight factors loaded on one construct. The values of the loadings ranged between .78 and .93. Thirdly, test of local independence of the test items was achieved by correlating the factors. A positive-low correlation ranging between .02 and .34 was observed between the factors. The final data collected were analyzed using test analysis modules (TAM) and differential item functioning for non-linear regression (difNLR) packages in open-source R computer program (R core team, 2015), version 3.4.3 via RStudio version 1.0.153. Uniform DIF was specifically adopted to detect if any difference in item difficulty existed between male and female Physics students. The TAM package is efficient in analyzing one-parameter logistic model. The partial credit model, PCM is a type of one-parameter logistic model. The discriminating parameter was constrained to a value of one. The difNLR package was used because PCM followed logistic or probit regression which was non-linear in nature. When the entire final data were first loaded at a time after invoking the packages with their dependencies, the PCM failed to converge. The data analyst continued to load the data in steps. The model converged for the first time when loaded item categories ranged from 1a to 6. When item category 7a was added, there was no model convergence. Item 7a was deleted and item 7b added. There was no model convergence. Other items were loaded, excluding items 7a and 7b. The model converged after 100 iterations, also known as computation cycles. Research questions were answered using frequencies, mean infit/outfit statistics, mean item difficulty and mean standard error. The null hypothesis was tested at .05 level of significance using Mantel-Haenszel's chi-square per degree of freedom probability value for the model. The modalities for goodness of infit/outfit statistics for logistic model ranges from .7 to 1.3 (Xiao, Han, Koenig, Xiong & Bao, 2018). Acceptable value of standard error of measurement used was less than .5 (Adonu, 2014). Acceptable item difficulty index ranged between -3 to 3 (Guyer & Thompson, 2011). For DIF, Guyer and Thompson (2011) observed that if the probability value associated to Chi-square per degree of freedom was less than .05, it indicated presence of DIF.

RESULTS:

The results are presented in line with the order of the research questions and hypothesis formulated for the study.



Research question one: What pattern of scoring was adopted by WAEC in 2018 WASSCE Physics essay marking scheme for Nigeria and The Gambia?

Data in Table1 shows that item1a was allocated 1 mark and the number of categories for obtainable marks was 2. Item1b was allocated 2 marks with number of categories of obtainable marks being 5. Items 2 and 3 were allocated 3 marks each with number of categories of obtainable marks being 4 each. Items 4a and 4b were respectively allocated 1 and 2 marks with number of categories of obtainable marks being 2 each.

Table1: Summary of the Pattern of scoring adopted by WAEC in 2018 WASSCE Physics essay marking scheme for Nigeria and The Gambia

S/N	Item	Maximum marks Allocated	Categories of obtainable marks	Number of categories	S/N	Item	Maximum marks Allocated	Categories of obtainable marks	Number of categories
1	1a	1	0,1	2	15	9a	4	0,1,2,3,4	5
2	1b	2	0,0.5,1, 1.5,2	5	16	9b	2	0,1,2	3
3	2	3	0,1,2,3	4	17	9c	2	0,2	2
4	3	3	0,1,2,3	4	18	9d	7	0,1,2,3,4,5,6,7	8
5	4a	1	0,1	2	19	10a	2	0,2	2
6	4b	2	0,2	2	20	10b	6	0,0.5,1,1.5,2,2 .5,3,3.5,4,4.5, 5,5.5,6	13
7	5	3	0,1,2,3	4	21	10c	7	0,0.5,1,1.5,2,2 .5,3,3.5,4,4.5, 5,5.5,6,6.5,7	15
8	6	3	0,0.5,1,1.5,2,2 .5,3	7	22	11a	4	0,2,4	3
9	7a	1.5	0, 1.5	2	23	11b	8	0,1,2,3,4,5,6,7 ,8	9
10	7b	1.5	0, 1.5	2	24	11c	3	0,1,2,3	4
11	8a	2	0,2	2	25	12a	2	0,2	2
12	8b	4	0,0.5,1,1.5,2,2 .5,3,3.5,4	9	26	12b	3	0,1,2,3	4
13	8c	6	0,0.5,1,1.5,2,2 .5,3,3.5,4,4.5, 5,5.5,6	13	27	12c	3	0,1,2,3	4
14	8d	3	0,1,2,3	4	28	12d	7	0,0.5,1,1.5,2,2 .5,3,3.5,4,4.5, 5,5.5,6,6.5,7	15

Item 5 had 3 and 4 as allocated marks and number of categories of obtainable marks respectively. Item 6 had 3 and 7 as allocated marks and number of categories of obtainable marks respectively. Items 7a and 7b had 1.5 marks and 2 categories each (fractional



polarization1). Items 8a, 8b, 8c and 8d were allocated 2, 4, 6 and 3 marks with categories 2, 9, 13 and 4 respectively. Items 9a, 9b, 9c and 9d were allocated 4, 2, 2 and 7 marks with categories 5, 3, 2 and 8 respectively. Items 10a, 10b and 10c were allocated 2, 6 and 7 marks with categories 2, 13 and 15 respectively. Items 11a, 11b and 11c were allocated 4, 8 and 3 marks with categories 3, 9 and 4 respectively. Items 12a, 12b, 12c and 12d were allocated 2, 3, 3 and 7 marks with categories 2, 4, 4 and 15 respectively. There were observed dichotomous scoring in items: 1a, 4a, 4b, 8a, 9c, 10a and 12a. However, fractional polarization scoring was observed for items 7a and 7b.

Research question two: What proportion of the item categories of WASSCE-2018-Physics-Essay test fits into partial credit model?

Data in Table 2 shows that items: 1a, 1b, 2, 3, 4a, 4b and 5 had mean infit and outfit statistic values as .77, 1.05, .87, .35, .68, .57, 1.11 and .88, .93, 1.21, 1.07, .77, 1.13, 1.20 respectively. Other items like 6, 8a, 8b, 8c, 8d had .63, .51, .97, 1.04, .69 and .64, .67, 1.03, 1.11 as infit and outfit values respectively. Items: 9a 9b, 9c and 9c had infit and outfit statistic whose values include .68, .78, .69, 1.07 and .94, .77, .61, .89 respectively. Items: 10a10b and 10c had infit values of magnitudes 1.02, .88 and .79 while outfit values have magnitudes .90, .74 and 1.07. Items: 11a, 11b, 11c and infit values of magnitudes .22, .60, .98 and outfit values of magnitudes .24, .68 and 1.12. Other items including items 12a, 12b, 12c and 12d had mean infit values of magnitudes .95, .73, .85 and .81 and mean outfit values of magnitudes.

Table2: Summary of the proportion of item categories of WASSCE 2018 Physics essay test that fits partial credit model

Item	Mean Infit statistic	Mean Outfit statistic	Item	Mean Infit statistic	Mean Outfit statistic
1a	0.77	0.88	9a	0.68	0.94
1b	1.05	0.93	9b	0.78	0.77
2	0.87	1.21	9c	0.69	0.61
3	0.35	1.07	9d	1.07	0.89
4a	0.68	0.77	10a	1.02	0.90
4b	0.57	1.13	10b	0.88	0.74
5	1.11	1.20	10c	0.79	1.07
6	0.63	0.64	11a	0.22	0.24
7a	-	-	11b	0.60	0.68
7b	-	-	11c	0.98	1.12
8a	0.51	0.67	12a	0.95	1.01
8b	0.97	0.95	12b	0.73	0.66
8c	1.04	1.03	12c	0.85	0.87
8d	0.69	1.11	12d	0.81	1.04
Grand Mean:				0.78	0.89



1.01, .66, .87 and 1.04 respectively. The grand mean for mean infit and outfit statistic were .78 and .87 respectively. Items 7a and 7b were not included in partial credit modeling because they contaminated the model. Item11a had a value below critical value. Thus, the proportion of the items that fit the partial credit model was 25/28 representing 89.3%. Appendix A shows the samples of item characteristic curve for three items (8c, 10a and 10c) that fit the partial credit model and one item (11c) that had a misfit.

Research question three: What are the values of difficulty-parameter with standard error of measurement in the test?

The result in Table 3 indicates that item 1a and 1b had mean difficulty levels with values -.43 (.01) and 1.24 (.12). The values indicated in bracket represent mean standard error associated to each difficulty parameter estimate. Items 2 and 3 had .35 (.05) and -.27 (.11).

Table 3: Summary of item difficulty parameter estimates with standard error

Item	Mean Difficulty level	Mean Standard Error	Item	Mean Difficulty level	Mean Standard Error
1a	-0.43	0.01	9a	0.79	0.06
1b	1.24	0.12	9b	-0.71	0.18
2	0.35	0.05	9c	1.04	0.18
3	-0.27	0.11	9d	0.71	0.09
4a	0.86	0.04	10a	1.13	0.16
4b	1.18	0.12	10b	1.21	0.06
5	1.71	0.13	10c	0.63	0.08
6	1.38	0.12	11a	-0.82	0.03
7a	-	-	11b	0.47	0.19
7b	-	-	11c	0.33	0.07
8a	-0.55	0.02	12a	0.96	0.01
8b	0.94	0.17	12b	0.83	0.10
8c	0.66	0.31	12c	1.14	0.14
8d	0.54	0.30	12d	0.89	0.19
Group Mean:				0.62	0.12

Items 4a, 4b, 5, 6, 8a, 8b, 8c and 8d had .86, 1.18, 1.71, 1.38, -.55, .94, .66 and .54 as mean item difficulty values and .04, .12, .13, .12, .02, .17, .31 as values of mean standard error respectively. In addition, items 9a, 9b, 9c, 9d, 10a, 10b and 10c had .79, -.71, 1.04, .71, 1.13, 1.21 and .63 as mean difficulty with .06, .18, .18, .09, .16, .06 and .08 as standard errors respectively. Items 11a, 11b, 11c, 11d, 12a, 12b, 12c and 12d had -.82, .47, .33, .96, .83, 1.14 and .89 as mean difficulty with .03, .19, .07, .01, .10, .14 and .19 as standard errors respectively. The overall mean difficulty with standard error was .62 (.12).



Hypothesis One:

The item categories of WASSCE-2018-Physics-Essay test do not exhibit uniform differential item functioning across students' gender.

The hypothesis sought to determine whether or not item difficulty varied across male (reference group) and female students (focal group). The result in Table 4 indicates that chi-square with its probability values for items: 1a, 1b, 2, 3, 4a, 4b, 5, 6, 7a, 7b, 8a, 8b, 8c and 8d included 2.079 (.13), 1.0032 (.23), .6775 (.78), 3.2104 (.44), 3.2471 (.22), 2.1439 (.07), 1.3876 (.13), 4.7392 (.08), 3.7682 (.67), .7291 (.83), 1.9835 (.69), .9186 (.93), .1942 (.57) and 1.1278 (.88) respectively.

Table 4: Summary of differential item functioning estimate

Item	Chi-square value	Probability (p) value	Item	Chi-square value	Probability (p) value
1a	2.0789	0.1302	9a	2.3032	0.7378
1b	1.0032	0.2311	9b	4.2287	0.8276
2	0.6775	0.7802	9c	3.8113	0.7823
3	3.2104	0.4391	9d	1.9823	0.7229
4a	3.2471	0.2236	10a	0.1178	0.0578
4b	2.1439	0.0704	10b	0.4638	0.1023
5	1.3876	0.1264	10c	1.2946	0.1183
6	4.7392	0.0826	11a	2.9453	0.7285
7a	3.7682	0.6691	11b	3.5921	0.8782
7b	0.7291	0.8267	11c	1.4982	0.6826
8a	1.9835	0.6892	12a	2.0091	0.6728
8b	0.9186	0.9321	12b	0.2228	0.7821
8c	0.1942	0.5729	12c	0.5653	0.8820
8d	1.1278	0.8824	12d	1.9981	0.2258

Items 9a, 9b, 9c, 9d, 10a, 10b, 10c, 11a, 11b, 11c, 12a, 12b, 12c, 12d had chi-square with probability values as 2.3032 (.74), 4.2287 (.83), 3.8113 (.78), 1.9823 (.72), .1178 (.06), .4638 (.10), 1.2946 (.12), 2.9453 (.73), 3.5921 (.88), 1.4982 (.68), 2.0091 (.67), .2228 (.78), .5653 (.88) and 1.9981 (.23) respectively. All the estimated probability values for chi-square values were within acceptable range, and showed presence of no DIF in any item. Item 6 in Appendix A shows a sample of uniform DIF plot for an item that did not exhibit uniform DIF.



DISCUSSION OF THE FINDINGS

The result of data analysis in Table 1 shows that 19 in 28 items, representing approximately 68% of the total items had categories of obtainable marks above three. However, 9 in 28 items representing approximately 32% of the total items were scored dichotomously. Fractional dichotomization scoring was observed in two out of the nine items. The partial credit model demands that an essay test should be partially scored (Primi & Primi, 2014). The allocated marks for an essay item should be broken up into at least three groups. Nineteen in twenty-eight items were deemed appropriately scored in line with Primi and Primi's observation. The marks allocated to the remaining nine items should be distributed into at least three groups. The process of distributing a mark allocated for an essay item to at least three groups eliminates fractional dichotomization of scores.

The result of data analysis in Table 2 shows that 25 in 28 items representing approximately 89% of Physics essay test fit into the partial credit model. A close look at Table 2 shows that two items that were fractionally dichotomized in scoring contaminated the dataframe. So they were excluded from the data before the model could run to identify one item as a misfit. Thus, it is adduced that fractional dichotomization scoring of the two items was the source of contamination to the dataframe. The items that fit the partial credit model were in tandem with indices of fit suggested by Adonu (2014) and Prama, Supahar, Warsono and Jumadi (2018) in their separate studies.

The result of data analysis in Table 3 shows that the standard error for the measurement of difficulty parameter low. The items of 2018 WASSCE Physics essay questions showed moderate item difficulty. The result is consistent with earlier finding by Adonu (2014) that WAEC sets moderate item difficulty questions in Physics.

The result of data analysis in Table 4 shows that none of the twenty-eight items of WASSCE 2018 Physics essay items exhibited differential item functioning in terms of the gender of the students.

EDUCATIONAL IMPLICATIONS

It can be deduced from the result of the study that the West African Examinations Council in its 2018 WASSCE Physics essay for Nigeria and The Gambia adopted dichotomized scoring in some items. This by implication shows that average ability students are put to a disadvantaged position in scoring such items correctly. High ability relative to low ability students are always put to a vantage position for essay items scored dichotomously. Thus the chances to score such items by low and middle ability students have been toughened by the scoring pattern.

It was observed that fractional dichotomization of the scoring pattern of two items in the test possibly accounted for the contamination of the dataframe for the partial credit model. The implication of this finding is that fractional dichotomization scoring should be eliminated by WAEC in its Physics marking scheme to make way for essay test evaluation, using the model



The difficulty parameter estimates in the test showed moderate difficulty. The implication of the finding is that WASSCE 2018 Physics essay was not too difficult for the students' level. Thus, evidence of poor achievement by students in the examination may not be attributed to item difficulty level.

The implication of the result which indicated that WASSCE 2018 Physics essay test did not function differently cross students' gender and had 68% of items that fit the partial credit model is that public confidence in WAEC conducted examinations should be sustained.

SUMMARY AND CONCLUSION

Secondary school students' poor achievement in WASSCE Physics has been observed as a factor of underdevelopment in Nigeria. This study was designed to evaluate the scoring pattern, item difficulty and differential item functioning in WASSCE 2018 Physics essay as a step towards increasing achievement in Physics. Based on the findings of the study, WAEC should not adopt dichotomization of scoring pattern in future Physics essay marking scheme. Ever allocated mark to Physics essay item should have at least three groups of obtainable marks. The WASSCE 2018 Physics essay items were moderately difficult. They did not function differentially across male and female Physics students' sub-groups.

REFERENCES

- Adonu, I. I. (2014). Psychometric Analysis of WAEC and NECO Practical Physics Test Using Partial Credit Model. *Unpublished Ph.D Thesis*, Department of Science Education, University of Nigeria, Nsukka.
- Anamezie, R.C. & Nnadi, F O. (2018). Parameterization of teacher-made Physics achievement test using deterministic-input-noisy-and-gate (DINA) model. *Journal of Education and Practice*, 9(32):101-109. www.iiste.org
- Federal Government of Nigeria,(2013). *National Policy on Education and a four-year strategic plan for the development of the Education sector*. Abuja, Nigeria: Government press.
- Guyer, R. & Thompson, N. (2011). Item response theory calibration software user manual for Xcalibre 4.2. Woodbury, MN 55125: USA, Assessment systems corporation.www.assessment.com
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174
- Matteucci, M. Stracqualursi, L. (2006). Students' Assessment via Graded Response Model. *Statistica*, Anno LXVI, (4): 435-447. Doi: 10.6092/issn1973-2201/1216
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16: 159-176

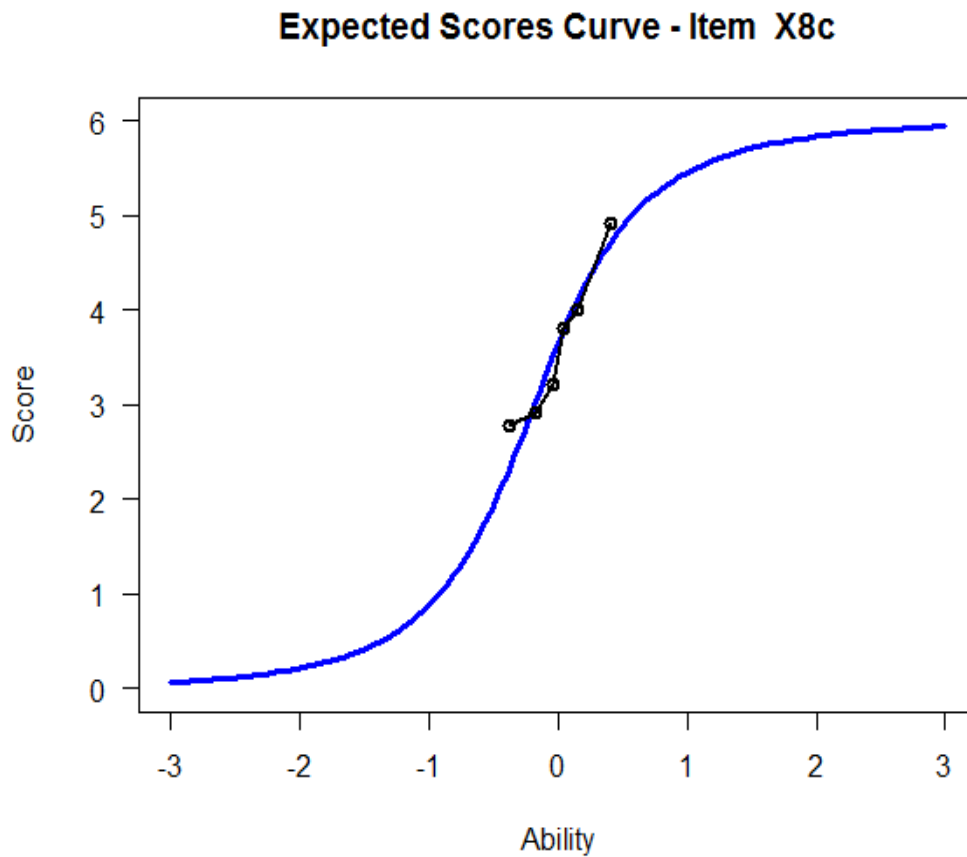


- Nworgu, B .G. (2016). Averting pedagogical failure in science: Insights from educational measurement and evaluation. 103rd Inaugural Lecture of the University of Nigeria, Nsukka delivered on Thursday, February 25.
- Nworgu, B. G. (2010). Challenges of Quality Assurance in a Challenging Global Economy. *Journal of Educational Assessment in Africa*, 5: 13-35
- Primi, T. C. N, & Primi, R. (2014). Rasch-Masters partial credit model in the assessment of children's' creativity in drawings. *The Spanish Journal of Psychology*, 17, e37:1-16. Doi: 10.1017/sjp.2014.36
- Qingpin, H. (2010). Maintaining Standards in on-demand Testing Using Item Response Theory
- R Core Team (2015). *R: A language and environment for statistical computing*. ISSN 3-9000051-07-0. R Foundation for statistical computing, Vienna, Austria. <http://www.R-project.org>
- Sung, H. J. & Kang, T. (2006). Choosing a polytomous IRT model using Bayesian Model selection methods. Paper presented at the National council on Measurement in Education in San-Francisco.
- Telima, A. (2018). Pupil Attainment in secondary school Physics: The case of Nigeria including implications for teachers and teacher education. *Journal of Social Science Research*, 12(2), 2783-2803. https://www.researchgate.net/publication/329366606_Pupil_Attainment_in_Secondary_School_Physics_The_Case_of_Nigeria_Including_Implications_for_Teachers_and_Teacher_Educators
- Thorpe, G. L. & Favia, A. (2012). Data Analysis Using Item Response Theory Methodology: An introduction to selected programs and applications.
- Tolu, O. O. & Olusegun, A. (2014). Effect of Enhanced explicit teaching strategy on pupils' learning outcome in primary schools. *Journal of the Science Teachers' Association of Nigeria (STAN)*, 49 (1), 1-18. www.stan.org.ng
- Ugodulunwa, C. A. (2015). Quality Assurance in Research, Assessment and Evaluation in Nigeria. *Nigerian Journal of Educational Research and Evaluation*, 14(1), 1-23
- Xiao,Y., Han,J., Koenig, K., Xiong, J. & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple choice test: A case-study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research*, 14, 1-18. https://www.researchgate.net/publication/326208567_Multilevel_Rasch_modeling_of_two-Tier_multiple_choice_test_A_case_study_using_Lawson's_classroom_test_of_scientific_reasoning



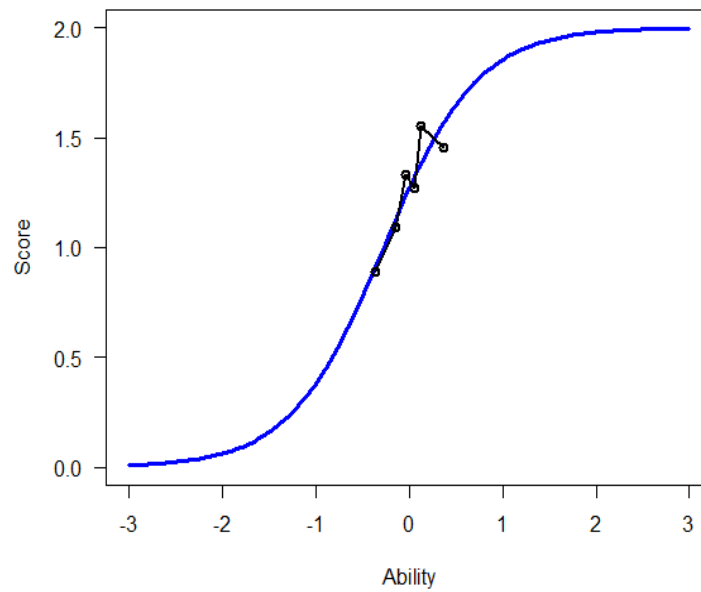
APPENDIX A:

Samples of item characteristic curves for three fit (items 8c, 10a & 10c) and one misfit (item11) items with profile plot for one non-DIF item (item6).

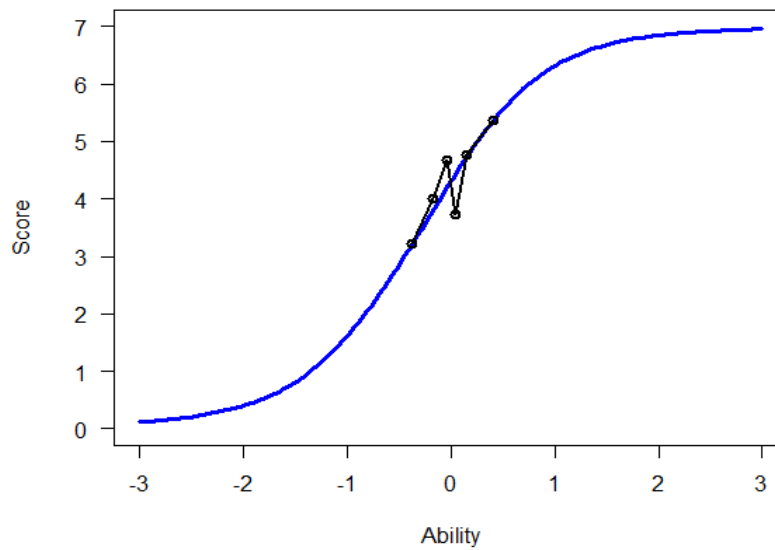




Expected Scores Curve - Item X10a

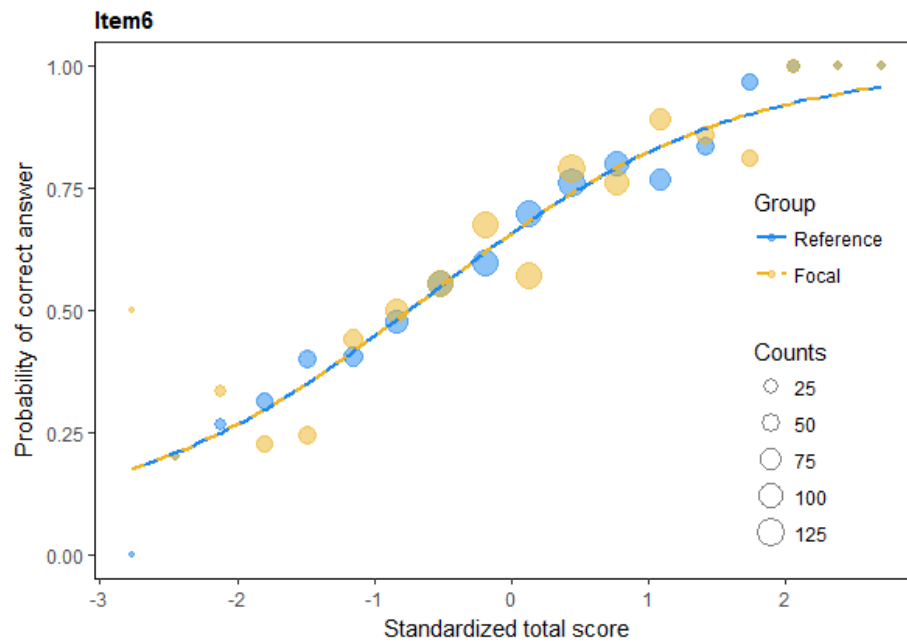
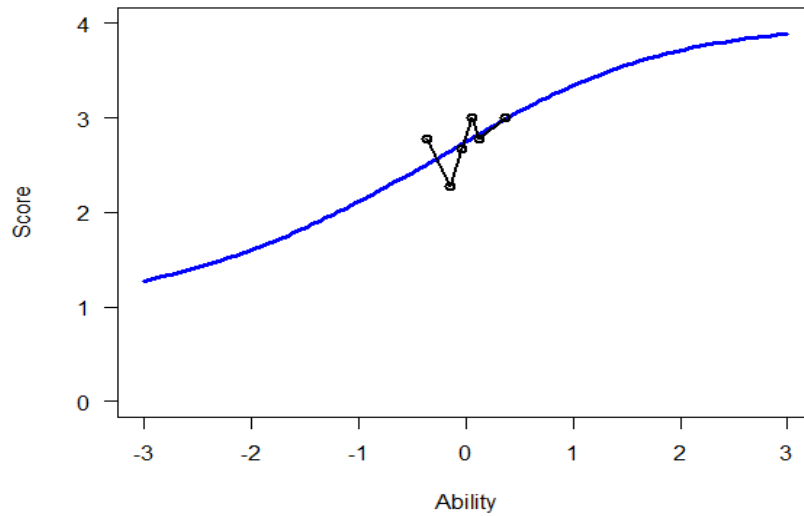


Expected Scores Curve - Item X10c





Expected Scores Curve - Item X11a





APPENDIX B:

R Codes used to evaluate the data:

```
library("TAM")
setwd("c:/users/NNADI/Documents/TAM")
D<-read.csv("PCM data.csv",header=TRUE, sep=",")
M<-tam(D , irtmodel ="1PL")
WLE1P<-tam.mml.wle2(M)# Extract Reliability
Ability<-M$theta #Extract person Ability
F<-tam.mml.fit(M) #Extract Item infit and outfit statistics
summary(M) #Extract item difficulty parameters and standard errors
plot(M,1:28) #Extract ICC plots for items 1-28
Prob<-IRT.irfprob(M) #Extract prob of theta
plot(Prob) #Extract IRF plot
h<-M$person# Extract individual scores and SD
msq.itemfit(M,1:28) # Extract mean square residual of items outfit and infit
IRT.modelfit(M) # Extract modelfit
cor(D, method = "pearson")#Correlation of item #to establish local independence
```

```
library("difNLR")
#import excel data: DIF.xlsx in Documents to global environment in Rstudio
#convert the data matrix to integer and the grouping column to
#character or numeric values using the following commands:
DIF$Item1a<-as.integer(DIF$Item1a)
DIF$Item1b<-as.integer(DIF$Item1b)
DIF$Item2<-as.integer(DIF$Item2)
DIF$Item3<-as.integer(DIF$Item3)
DIF$Item4a<-as.integer(DIF$Item4a)
DIF$Item4b<-as.integer(DIF$Item4b)
DIF$Item5<-as.integer(DIF$Item5)
DIF$Item6<-as.integer(DIF$Item6)
DIF$Item7a<-as.integer(DIF$Item7a)
DIF$Item7b<-as.integer(DIF$Item7b)
DIF$Item8a<-as.integer(DIF$Item8a)
DIF$Item8b<-as.integer(DIF$Item8b)
DIF$Item8c<-as.integer(DIF$Item8c)
DIF$Item8d<-as.integer(DIF$Item8d)
DIF$Item9a<-as.integer(DIF$Item9a)
DIF$Item9b<-as.integer(DIF$Item9b)
DIF$Item9c<-as.integer(DIF$Item9c)
```



```
DIF$Item9d<-as.integer(DIF$Item9d)
DIF$Item10a<-as.integer(DIF$Item10a)
DIF$Item10b<-as.integer(DIF$Item10b)
DIF$Item10c<-as.integer(DIF$Item10c)
DIF$Item11a<-as.integer(DIF$Item11a)
DIF$Item11b<-as.integer(DIF$Item11b)
DIF$Item11c<-as.integer(DIF$Item11c)
DIF$Item12a<-as.integer(DIF$Item12a)
DIF$Item12b<-as.integer(DIF$Item12b)
DIF$Item12c<-as.integer(DIF$Item12c)
DIF$Item12d<-as.integer(DIF$Item12d)
DIF$gender<-as.character(DIF$gender)
group=DIF$gender
str(DIF)
Data<-DIF[,1:28]
str(Data)
y<-difNLR(Data, group, focal.name="1",
  model="1PL",constraints = "a", type = "udif",
  match = "zscore",anchor = NULL, purify = FALSE,
  nrIter = 1000, test = "LR",
  alpha = 0.05, p.adjust.method = "bonferroni")
plot(y)
coef(y)
```